NEW YORK UNIVERSITY
INSTITUTE OF MATHEMATICAL SCIENCES
LIBRARY
25 Waverly Place, New York 3, N. Y.

IMM-NYU 278
NOVEMBER 1960

NEW YORK UNIVERSITY
INSTITUTE OF
MATHEMATICAL SCIENCES

Maximum Likelihood Characterization of Distributions

HENRY TEICHER

PREPARED UNDER
CONTRACT NO. NONR-285(38)
OFFICE OF NAVAL RESEARCH
AND
CONTRACT NSF G-6331
NATIONAL SCIENCE FOUNDATION

REPRODUCTION IN WHOLE OR IN PART
IS PERMITTED FOR ANY PURPOSE
OF THE UNITED STATES GOVERNMENT.

IMM-278
c.1

IMM-NYU 278
November 1960

New York University
Institute of Mathematical Sciences

MAXIMUM LIKELIHOOD CHARACTERIZATION OF DISTRIBUTIONS

Henry Teicher

New York University
and
Purdue University

This report represents results obtained at the Institute of Mathematical Sciences, New York University, under the sponsorship of Contract No. Nonr-285(38), with the Office of Naval Research and Contract NSF G-6331, National Science Foundation.

1960

1. Introduction. It is a commonplace observation that the sample mean and sample variance from a normal population (based on a random sample) are stochastically independent. Considerably less prosaic is the converse proposition first proved in 1936 by R. C. Geary [2] (under superfluous restrictions) to the effect that the independence of these two statistics entails normality of the underlying population. This, plus the theorem that if two linear combinations (non-zero coefficients) of a pair of independent random variables are themselves independent, the variables are normally distributed which was proved by Kac in 1939, [4] are harbingers of what are today referred to as characterization theorems. An extensive bibliography of such theorems appears in [5]. Most of these results have the format: if such-and-such statistics are independent (alternatively, if the distribution of such-and-such statistic is thus-and-so), the underlying population is so-and-so.

The ensuing theorems belong to this genre but adopt a maximum likelihood posture. The first deals with translation (location) parameter and the others with scale parameter families. The former is seen to be a generalization and modernization of a result of Gauss [7] when the latter is suitably interpreted and rescued from its least squares context.

Theorem 1: Let $F(x - \theta)$, θ real, be a translation parameter family of absolutely continuous distributions on the real line. If, for all (random) samples of size two and three, a maximum likelihood estimator of θ is the sample arithmetic mean, then $F(x)$ is the normal distribution.

Proof: Let $f(x) = F'(x) \geq 0$ and $h(x) = \log_e f(x)$ where $h(x)$ may assume the extended real values $\pm \infty$. Recall that a maximum likelihood estimator based on a random sample $x_1, x_2, \ldots x_N$ of observations from $F(x - \theta)$ is a value of θ which maximizes $\prod_{i=1}^{N} f(x_i - \theta)$.

Thus, if $\bar{x} = \frac{1}{n+1} \sum_{i=1}^{n+1} x_i$, it follows from the hypothesis that for $n = 1, 2$, and all real $x_1, x_2, \ldots, x_{n+1}$ and θ,

$$\sum_{i=1}^{n+1} h(x_i - \bar{x}) \geq \sum_{i=1}^{n+1} h(x_i - \theta)$$

or

$$\sum_{i=1}^{n+1} h(y_i) \geq \sum_{i=1}^{n+1} h(y_i - \theta) \quad \text{when} \sum_{i=1}^{n+1} y_i = 0;$$

that is, for $n = 1,2$ and all real $y_1, y_2, \ldots, y_n$, θ

$$(1) \quad \sum_{i=1}^{n} h(y_i) + h(-\sum_{i=1}^{n} y_i) \geq \sum_{i=1}^{n} h(y_i - \theta) + h(-\sum_{i=1}^{n} y_i - \theta) .$$

As will be seen, under the meager assumptions inherent in the probabilistic framework viz., $f(x)$ measurable and positive on a set of positive lebesgue measure, the functional inequality (1) determines $h(x)$ (and therefore $f(x)$).

In particular, (1) implies

$$(2) \qquad nh(y) + h(-ny) \geq nh(y - \theta) + h(-ny - \theta)$$

which, for $n = 1$, becomes

$$(3) \qquad 0 \geq h(-y-\theta) - h(y) - h(-y) + h(y-\theta).$$

Note that if in (1), θ is replaced by $-\theta$ and y_i by $-y_i$ the resulting inequality when added to (1) reveals that if $g(y)$ is a solution of (1), so is $h(y) = g(y) + g(-y)$. Thus, any solution has a symmetric counterpart and we therefore confine attention at first to symmetric solutions h of (1), which then takes the form

$$(1)' \quad \sum_{i=1}^{n} h(y_i) + h\left(\sum_{i=1}^{n} y_i\right) \geq \sum_{i=1}^{n} h(y_i - \theta) + h\left(\sum_{i=1}^{n} y_i + \theta\right).$$

Similarly, (3) becomes

$$(3)' \qquad 0 \geq h(y + \theta) - 2h(y) + h(y - \theta)$$

for all y, θ and asserts the concavity property for any symmetric solution h of (1). Since the density function $f(x)$ is necessarily measurable, $h(x)$ is likewise, whence [1,6] any solution h of (1)' is a continuous concave function.

Set $n = 1$, $y = 0$ in (2) obtaining $h(0) \geq h(-\theta)$, all θ; then $h(0) \neq -\infty$ since otherwise $f(x) \equiv 0$, an impermissible state of affairs; further, by concavity, $h(x)$ is finite at $x = 0$ and hence also in an interval containing the origin. If there exists c, $0 < c < \infty$ with $h(c) = -\infty$, let c' be the infimum of such values c. Then taking $n = 2$, $y_1 = y_2 = c'/2$, $\theta = -c'/4$ in (1)' yields $2h(c'/2) + h(c') \geq h(3c'/4)$ which is manifestly impossible. Thus, $h(x)$ is finite for all x.

Let D denote the complement of the (at most countable) set of points at which $h(x)$ is not differentiable and denote by $q(x)$ the derivative of $h(x)$. Then $q(x)$ is monotone and defined for all $x \in D$.

The fact, as expressed by (1)', that $\theta = 0$ maximizes $\sum_{i=1}^{n} h(y_i - \theta) + h(\sum_{1} y_i + \theta)$ now requires that

$$(4)' \qquad -\sum_{i=1}^{n} q(y_i) + q(\sum_{i=1}^{n} y_i) = 0$$

for all $y_i \in D$ such that $\sum_{i=1}^{n} y_i \in D$, $n = 1,2$. For $n = 2$, this becomes

$$(5)' \quad q(y_1) + q(y_2) = q(y_1 + y_2) \quad \text{for } y_1, y_2, y_1 + y_2 \in D.$$

Since the only monotone (measurable) solution of Cauchy's functional equation is $q(y) = c_1 y$, it follows that the only (measurable) symmetric solution of (1) is $h(y) = -cy^2 + d$ for $y \in D$ and therefore, by continuity, for all real y.

Suppose next that $g(y)$ is any (measurable) solution of (1). Then, necessarily for some real c and d, $g(y) = \frac{-cy^2+d}{2} + b(y)$ where $b(y)$ is an odd function. For, by the preceding $g(y) + g(-y) = -cy^2 + d$ for some c,d and this implies that $b(y) = g(y) + \frac{cy^2-d}{2}$ satisfies $b(y) = -b(-y)$.

Substituting $g(y) = b(y) + \frac{d-cy^2}{2}$ in (3) yields

$$(4) \qquad c\theta^2 \geq b(y - \theta) - b(y + \theta).$$

Replacing y by $-y$ and θ by $-\theta$ in (4) and combining the result with (4), there follows

$$(5) \qquad |b(y + \theta) - b(y - \theta)| \leq c\theta^2$$

which in turn, necessitates $c \geq 0$ and implies for every positive integer k and all real y, θ that

$$(6) \qquad |b(y + \theta/k) - b(y)| \leq \frac{c\theta^2}{4k^2} .$$

Thus, for every positive integer k, and all real y and θ,

$$|b(y + \theta) - b(y)| = \left|\sum_{i=1}^{k}[b(y + \frac{i\theta}{k})-b(y+ \frac{(i-1)\theta}{k})]\right| \leq \frac{c\theta^2}{4k}$$

whence $b(y + \theta) \equiv b(y) \equiv b(0) = 0$.

Consequently, the only measurable solution of (1) is $h(x) = -cx^2/2 + d$, $c \geq 0$ and thus the only density function satisfying the conditions of the theorem is $f(x) = ae^{-cx^2/2}$ for all real $x(a > 0, c > 0)$, that is, the normal density function.

Remark 1: The integers two and three of the theorem may clearly be replaced by other pairs. It seems most desirable, however, to state the result with minimal values of n.

Remark 2: If the density f is an even function, the translation parameter θ is a median of the distribution and also the mean if this latter exists. In such cases (excluding the normal), the theorem implies that the sample mean is not a maximum likelihood estimator of the population mean. This is readily seen, for example, if

$$f(x,\theta) = C_\alpha \exp\left\{-|x-\theta|^\alpha\right\}, \quad \alpha \neq 2.$$

Remark 3: It seems of interest to note in the case where $F(x-\theta)$ is a rectangular distribution with mean θ, i.e. $F'(x) = 1$, $|x-\theta| \leq {}^1/2$ that while $\bar{x}$ is a maximum likelihood estimator of θ for all random samples of size two, it is not a maximum likelihood estimator of θ for <u>all</u> random samples of size three.

2. <u>Scale parameter families</u>. Consider[1] an absolutely continuous scale parameter family of distributions $\mathcal{F} = \{F(x/\sigma), \sigma > 0\}$. The joint density function of independent random variables $x_1 \ldots x_n$ each distributed as $F(x/\sigma)$ is

$$\sigma^{-n} \prod_{i=1}^{n} f(x_i/\sigma)$$

where $f(x) = F'(x)$. Let $h(x) = \log_e f(x)$ where h may assume the extended real values $\pm\infty$. If $\hat{\sigma} = \hat{\sigma}(x_1 \ldots x_n)$ denotes a maximum likelihood estimator of σ,

$$\sum_{i=1}^{n} h\left(\frac{x_i}{\hat{\sigma}}\right) - n \log_e \hat{\sigma} \geq \sum_{i=1}^{n} h\left(\frac{x_i}{\sigma}\right) - n \log_e \sigma .$$

Let $y_i = x_i/\hat{\sigma}$, $\lambda = \hat{\sigma}/\sigma$. Then, if $\hat{\sigma}$ is a homogeneous function of degree one in $x_1, x_2, \ldots x_n$, the preceding becomes

$$\frac{1}{n} \sum_{i=1}^{n} [h(y_i) - h(\lambda y_i)] \geq \log_e \lambda \tag{7}$$

for all $\lambda > 0$ and $y_1, \ldots y_n$ satisfying

$$\hat{\sigma}(y_1, \ldots y_n) = 1. \tag{8}$$

Inspection shows that $h(y) = -\log_e y + \text{const.}$ satisfies (7) with equality holding for all choices of $y_1 \ldots y_n$ and a fortiori for y's satisfying (8). However $f(y) = c/y$ is not integrable on $(0,\infty)$. Further, f cannot be truncated to a finite interval since once $f(y) = 0$ for $y > 0$, f, if continuous, is identically zero for $y > 0$.

In the following theorems the indispensable absolute continuity assumption is augmented by a possibly dispensable continuity assumption of the density. The seemingly ad hoc condition (ii) on the other hand appears to be crucial.

Theorem 2: Let $\{F(x/\sigma), \sigma > 0, F(0) = 0\}$ be an absolutely continuous scale parameter family of c.d.f.'s satisfying

(i) $F'(x)$ continuous on $(0,\infty)$

(ii)[2] $\lim_{y \searrow 0} \frac{F'(y)}{F'(\lambda y)} = 1$, all $\lambda > 0$.

If, for all random samples, a maximum likelihood estimator of σ is the sample arithmetic mean, then F is the exponential distribution, i.e., $f(x) = e^{-x}, \; x \geq 0$

$0 \quad , \; x < 0.$

Proof: From prior remarks, (7) obtains with (8) becoming

$$(8.1) \qquad \sum_{i=1}^{n} y_i = n, \quad y_i > 0, \quad i = 1 \ldots n.$$

Take $y_i = k/m$, $1 \leq i \leq m$; $y_i = \frac{n-k}{n-m}$, $m + 1 \leq i \leq n$ in (7), obtaining

$$\frac{m}{n}[h(k/m) - h(\frac{\lambda k}{m})] + \frac{n-m}{n}[h(\frac{n-k}{n-m}) - h(\frac{\lambda(n-k)}{n-m})] \geq \log_e \lambda.$$

Let $\frac{k}{m} \longrightarrow \alpha$, $\frac{m}{n} \longrightarrow c$ where $0 < c, \alpha < 1$. Then for all $\lambda > 0$

$$(9) \quad c[h(\alpha) - h(\lambda\alpha)] + (1-c)[h(\frac{1-\alpha c}{1-c}) - h(\frac{\lambda(1-\alpha c)}{1-c})] \geq \log_e \lambda.$$

Let $\alpha \longrightarrow 0$ and set $y = (1-c)^{-1}$ in (9) yielding via (ii)

$$(10) \quad y^{-1}[h(y) - h(\lambda y)] \geq \log_e \lambda, \quad y \geq 1, \lambda > 0.$$

Replace λ by λ^{-1} in (10) and combine the result with (10) to obtain

$$(11) \quad 0 \geq h(\lambda y) - 2h(y) + h(y/\lambda), \quad y \geq 1, \lambda > 0.$$

This asserts that $H(y) = h(e^y)$ is concave for $y \geq 0$ and hence that $h(y)$ is differentiable in $(1,\infty)$ except perhaps on a countable subset D thereof. From (10), for $\lambda < 1$ and $y \geq 1$

$$\frac{h(\lambda y)-h(y)}{y(\lambda-1)} \geq \frac{\log_e \lambda}{1-\lambda} \geq \frac{h(y/\lambda)-h(y)}{\lambda y[(1/\lambda)-1]}$$

whence $(\lambda \nearrow 1)$, $h'(y) = -1$ on $(1,\infty) - D$. Then by continuity,

$$(12) \quad h(y) = -y + c, \quad y \geq 1.$$

Next, choosing $y_i > 1$, $i = 1,\ldots, r < n$ and $y_j > 1$, $j > r$ in (7) and employing (12) and (8.1), we find for all $\lambda > 0$, $r < n$ and y_i in $(0,1)$ that

$$(13) \quad \sum_{i=1}^{r} [h(y_i) - h(\lambda y_i) + (1-\lambda)y_i] \geq n[\log_e \lambda + 1 - \lambda].$$

For $0 < y < 1$, (13) asserts $(r=1)$ that

$$(14) \qquad \frac{1}{n}[h(y) - h(\lambda y)] \geq \log_e \lambda + (1-\lambda)(1-{}^{y}/_{n}).$$

But for $0 < x < 1$, $\log_e\lambda + x(\lambda-1) > 0$ for all λ sufficiently close to and larger than unity. Thus, from (14), h is monotone decreasing in $(0,1)$.

Set $h = h_1 + h_2$ where h_2 is absolutely continuous and h_1 is singular, i.e., $h_1'(y) = 0$ a.e. in $(0,1)$ and $h_1(y) \equiv 0$ for $y > 1$.

Again, taking $r = 1$ in (13), there follows for $0 < y < 1$ and $\lambda < 1$

$$(15) \qquad \frac{h(\lambda y)-h(y)+(\lambda-1)y}{\lambda-1} \geq n\left[\frac{\log_e \lambda}{1-\lambda} + 1\right].$$

This implies $y[h_2'(y) + 1] \geq 0$ almost everywhere in $(0,1)$. Similarly, replacing λ by λ^{-1} in (13), the inequality is reversed. Hence, $h_2'(y) = -1$ almost everywhere in $(0,1)$ implying $h_2(y)=-y+c_1$. By continuity $c = c_1$.

Utilizing this result in (15), there follows

$$(16) \qquad \overline{\lim_{\lambda\to 1}} \frac{h_1(\lambda y)-h_1(y)}{y(\lambda-1)} \geq \overline{\lim_{\lambda\nearrow 1}} \frac{h_1(\lambda y)-h_1(y)}{y(\lambda-1)} \geq 0$$

for all y in $(0,1)$. Hence $h_1(y) \equiv 0$.

Thus, $f(y) = ae^{-y}$ in $(0,\infty)$. Finally, $\int_0^\infty f(y)dy = 1$ requires that $a = 1$.

Theorem 3: Let $\{F(x/\sigma), \sigma > 0\}$ be an absolutely continuous scale parameter family of c.d.f.'s satisfying

(i) $F'(x)$ continuous on $(-\infty, \infty)$

(ii)^2 $\lim_{y \to 0} \frac{F'(y)}{F'(\lambda y)} = 1$, all $\lambda > 0$.

If, for all random samples, a maximum likelihood estimator of σ is $(\frac{1}{n}\sum_{i=1}^{n} x_i^2)^{1/2}$, then F is the normal distribution, i.e., $f(x) = F'(x) = \frac{1}{\sqrt{2\pi}} \; -x^2/2$.

Proof: Here, (8) specializes to

$$\sum_{i=1}^{n} y_i^2 = n. \tag{8.2}$$

Setting $y_i = \pm\sqrt{k/m}, 1 \leq i \leq m$; $y_i = \pm\sqrt{\frac{n-k}{n-m}}$, $m + 1 \leq i \leq n$, we find analogous to (9) that

$$c[h(\alpha) - h(\lambda\alpha)] + (1-c)[h(\pm\sqrt{\frac{1-\alpha^2 c}{1-c}}) - h(\pm\lambda\sqrt{\frac{1-\alpha^2 c}{1-c}})] \geq \log_e \lambda \; ; \tag{9}'$$

here, however, $|\alpha| \leq 1$. Thus,

$$y^{-2}[h(y) - h(\lambda y)] \geq \log_e \lambda, \quad |y| \geq 1, \quad \lambda > 0. \tag{10}'$$

Again, h is differentiable, this time in the region A: $|y| \geq 1$, except perhaps on a countable subset D' thereof. Proceeding as in the proof of Theorem 2, we find that $h'(y) = -y$ for y in A - D' and hence that $h(y) = -y^2/2 + c$ for $|y| \geq 1$. The analogue of (13) is

$$\sum_{i=1}^{r} [h(y_i) - h(\lambda y_i) + (1-\lambda^2)y_i^2/2] \geq n[\log_e\lambda + \frac{1-\lambda^2}{2}] \tag{13}'$$

where $|y_i| < 1$ and $r < n$. This implies that h is decreasing in (0,1) and increasing in (-1,0). An argument paralleling that of Theorem 2 yields $h(y) = -y^2/2 + c_1$ for $|y| < 1$ implying $f(y) = ae^{-y^2/2}$, all y. Finally, $a = 1/\sqrt{2\pi}$.

References

1. H. Blumberg, "On Convex Functions" Trans. Amer. Math. Soc. vol. 20(1919) pp. 40-44.

2. R. C. Geary, "Distribution of 'Student's' ratio for non-normal samples", Jour. Roy. Stat. Soc. Suppl. Ser. B, vol. 3 (1936) pp. 178-184.

3. S.G. Ghurye, "Characterization of Some Location and Scale Parameter Families of Distributions", Contributions to Probability and Statistics: Essays in honor of Harold Hotelling, Stanford University Press, Stanford,Calif. 1960.

4. M. Kac, "On a characterization of the normal distribution", Amer. Jour. Math., vol. 61(1939) pp. 726-728.

5. E. Lukacs, "Characterization of Populations by Properties of Suitable Statistics", Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley and Los Angeles, 1956, pp. 195-214.

6. W. Sierpinski, "Sur les fonctions convexes mesurables", Fundamenta Math.1 (1920) pp. 125-9.

7. Carl Friedrich Gauss, Werke, Theoria Motus Corporum Coelestium, Liber II, Sectio III, pp. 240-244. See also "Gauss's Work on the Theory of Least Squares", by Hale F. Trotter, pp. 127-133, Statistical Techniques Research Group, Princeton University, Princeton, N.J.

Footnotes

1. The standard device of reducing a scale parameter family to a translation parameter family (so as to utilize Theorem 1) appears fruitless here.

2. This condition is automatically fulfilled if the right-hand derivative at zero satisfies $0 < F'(0) < \infty$.

NEW YORK UNIVERSITY
INSTITUTE OF MATHEMATICAL SCIENCES
LIBRARY
25 Waverly Place, New York 3, N. Y.

FEB 23 1961

Date Due

PRINTED IN U. S. A.

NYU c.1
IMM-278
Teicher

Maximum likelihood character-
ization...

NYU c.1
IMM-278
Teicher
AUTHOR
Maximum likelihood
TITLE
characterization...

DATE DUE	BORROWER'S NAME	ROOM NUMBER

N. Y. U. Institute of
Mathematical Sciences
25 Waverly Place
New York 3, N. Y.